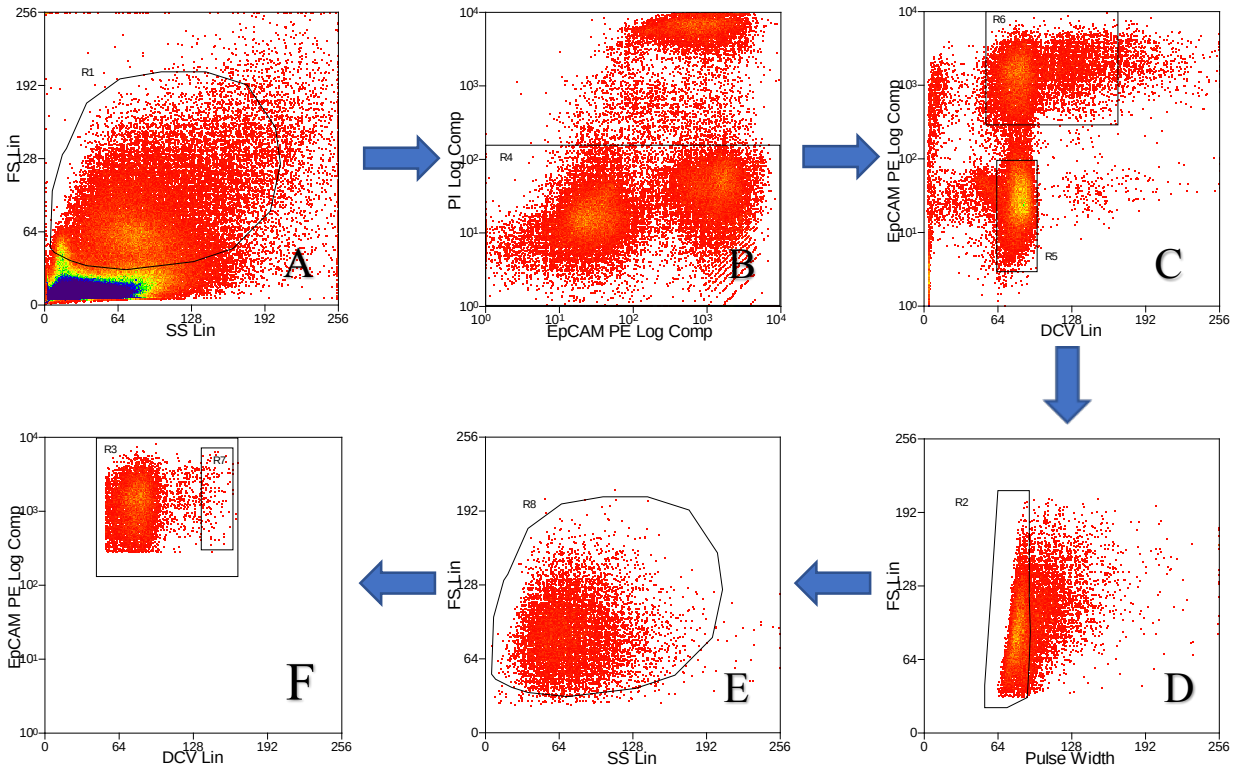
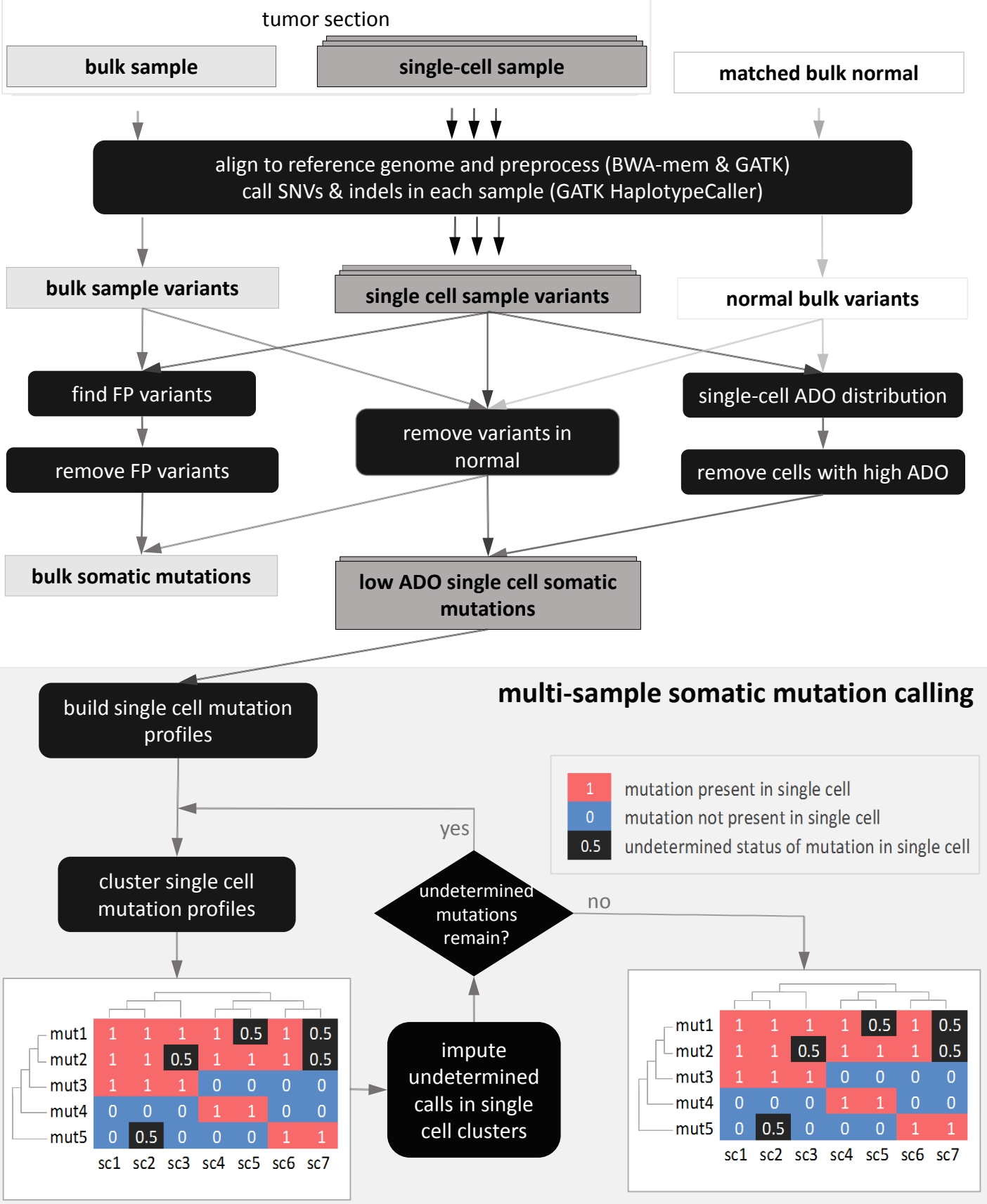
**Single-cell sequencing defines genetic heterogeneity in pancreatic cancer precursor lesions**
Kuboki, Fischer, Beleva Guthrie *et al*. *J Pathol* 2018 (DOI: 10.1002/path.???? copy ed please add number)

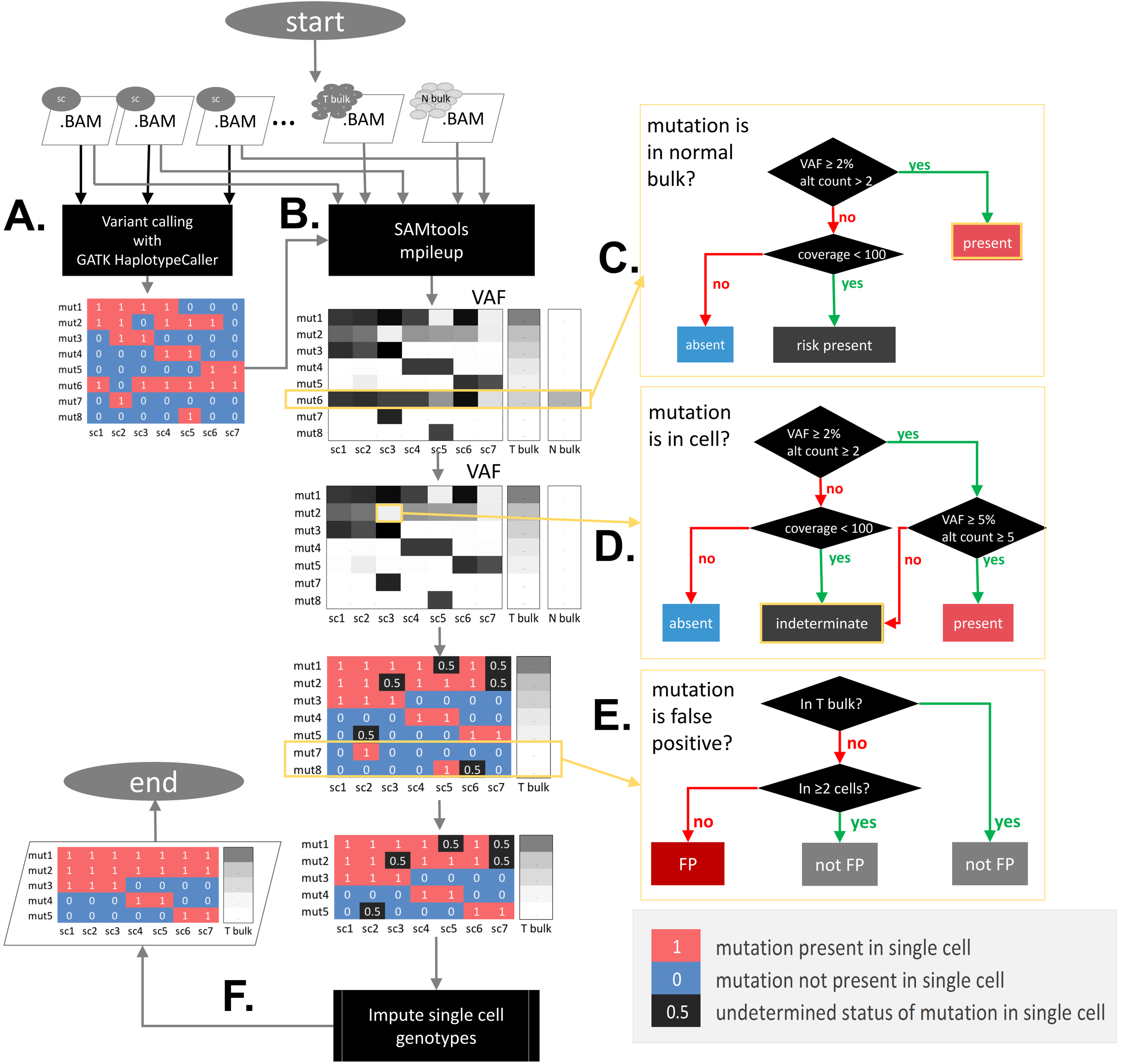
Supplementary Figures S1–S6



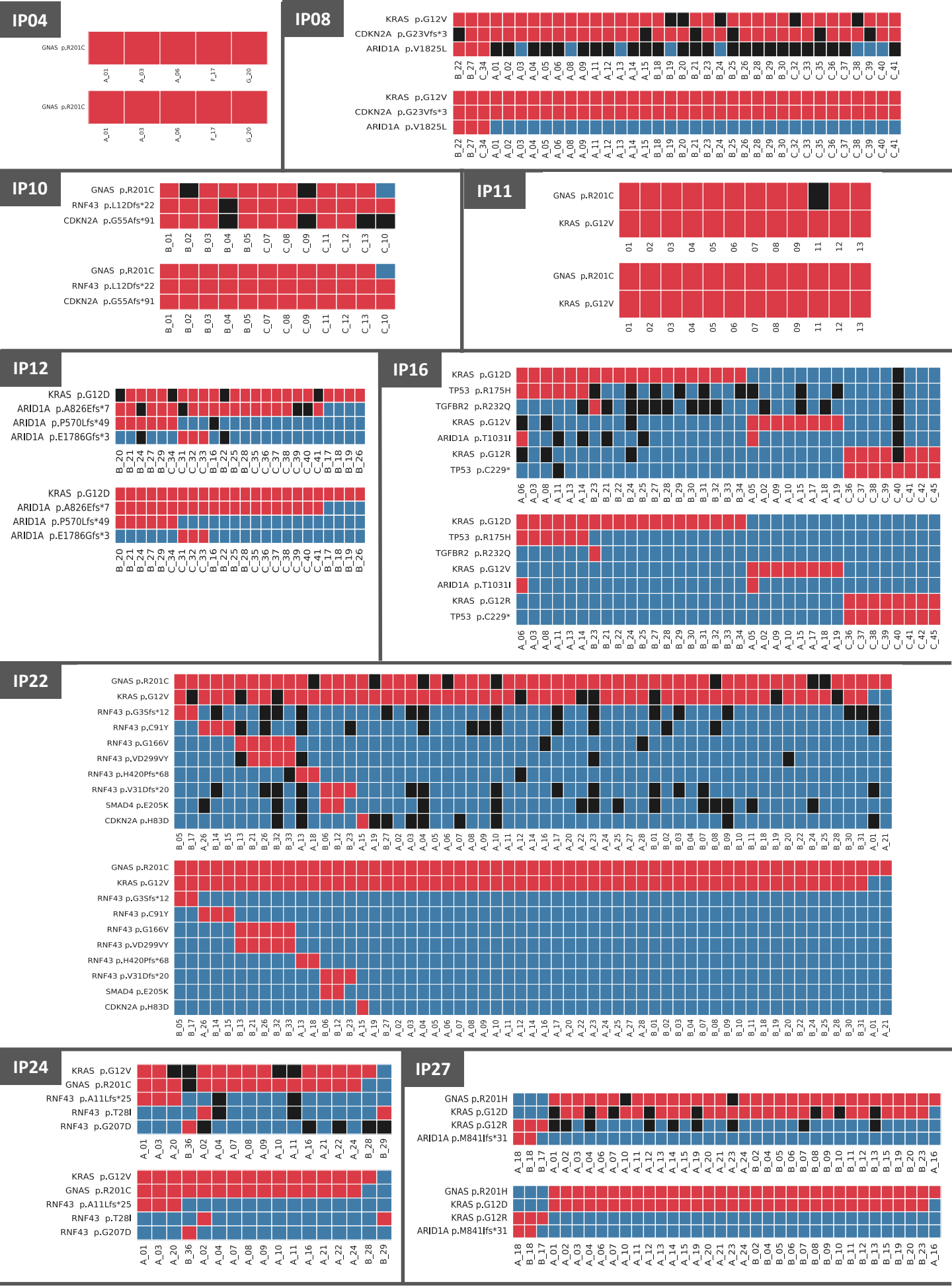

**Supplementary Figure S1. Gating scheme used for single cell sorting by fluorescence activated cell sorting (FACS).** On the scatter plot, events in region with morphologically intact cells dissociated from IPMN tissue (dotplot A) were gated on live cells (PI-, dotplot B). G1 and G2/M phase epithelial cells (EpCAM+) were identified using the DyeCycle Violet (DCV) DNA dye (dotplot C), and further gated to solely identify epithelial singlets (dotplot D). EpCAM+/PI-/G1 and G2 singlets were backgated on forward and side scatter to ensure that the epithelial cell recovery from the light scatter gate was >95% (dotplot E) and the bulk cells were sorted as EpCAM+DCV+ (dotplot F) with DNA content from G1 to G2.



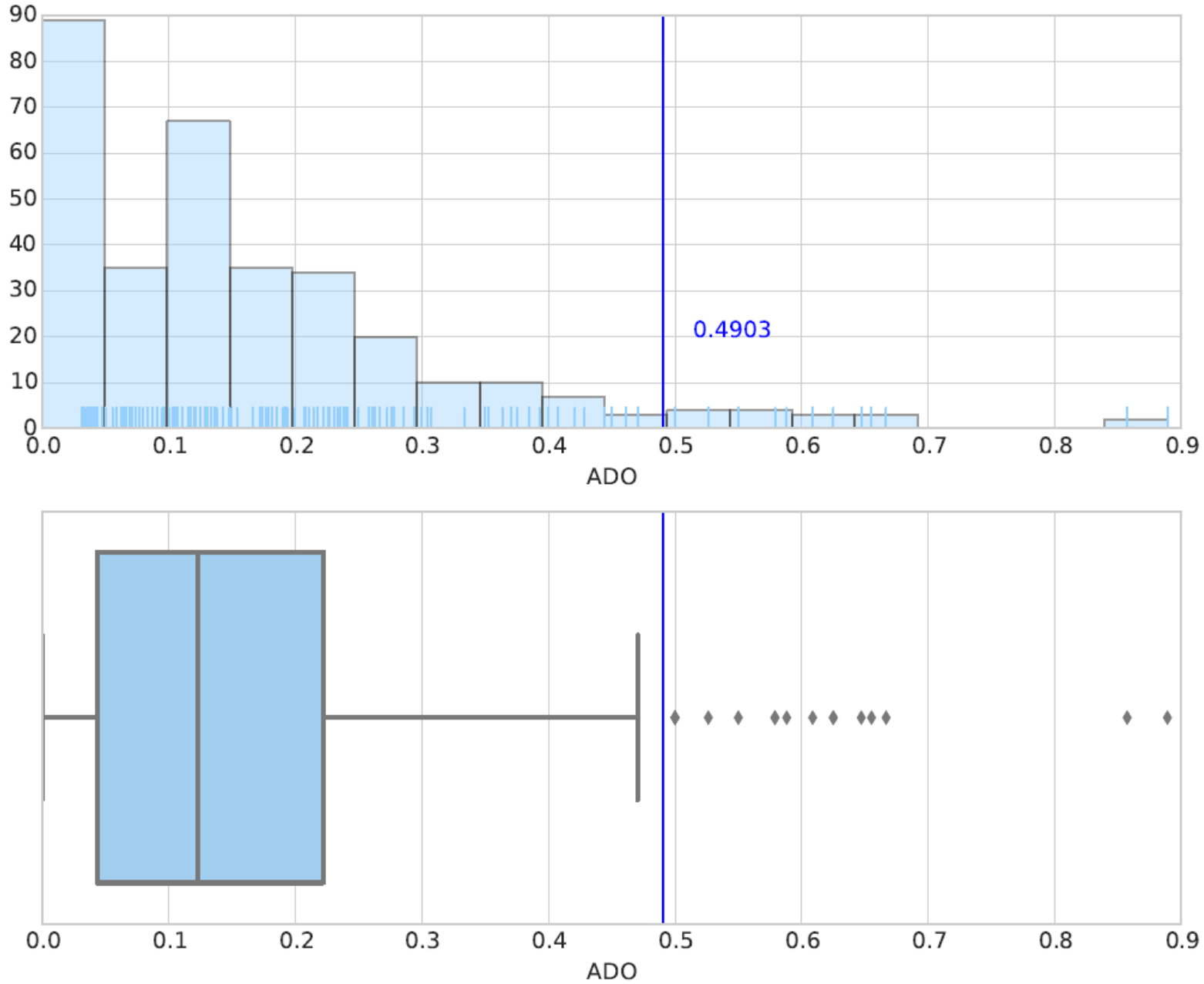

**Supplementary Figure S2. Computational workflow for mutation calling and error estimation.** For each of the cases analyzed, the input to our mutation calling pipeline was the set of next generation sequencing reads from tumor single-cell and bulk samples, and from the matched normal bulk sample. After aligning the reads to the human reference genome, non-reference variants (SNVs and indels) were called in each sample independently. Two filters were applied to the variants detected in the tumor samples. First, variants called in only one single cell and not present in the bulk tumor were removed as likely false positives (FPs), and then, variants called in the normal were subtracted from the variants called in tumor samples, such that only tumor somatic variants remain. The normal bulk heterozygous variants were used to estimate the allelic drop-out (ADO) rate in each single cell, which was defined here as the proportion of normal heterozygous variants that were homozygous reference or alternate in the single cell. Through analysis of the ADO rate distribution of the single cells from all IPMN cases, we established a maximum ADO rate threshold for including single cells in further analysis. From the set of filtered somatic variants in single cells passing the ADO rate threshold, we constructed single-cell mutation profile matrices, in which rows were mutations and columns were single-cell samples. The mutation profile matrices contained indeterminate elements, assigned to values of 0.5, due to insufficient sequencing information for definitively calling or rejecting the mutation in that single cell sample (for example, low but nonzero variant allele frequency or read count, or low coverage at the site). Our iterative procedure clustered single cell samples and mutations and imputed undetermined mutation status in a cell from a cluster of the most genetically similar cells in the tumor.



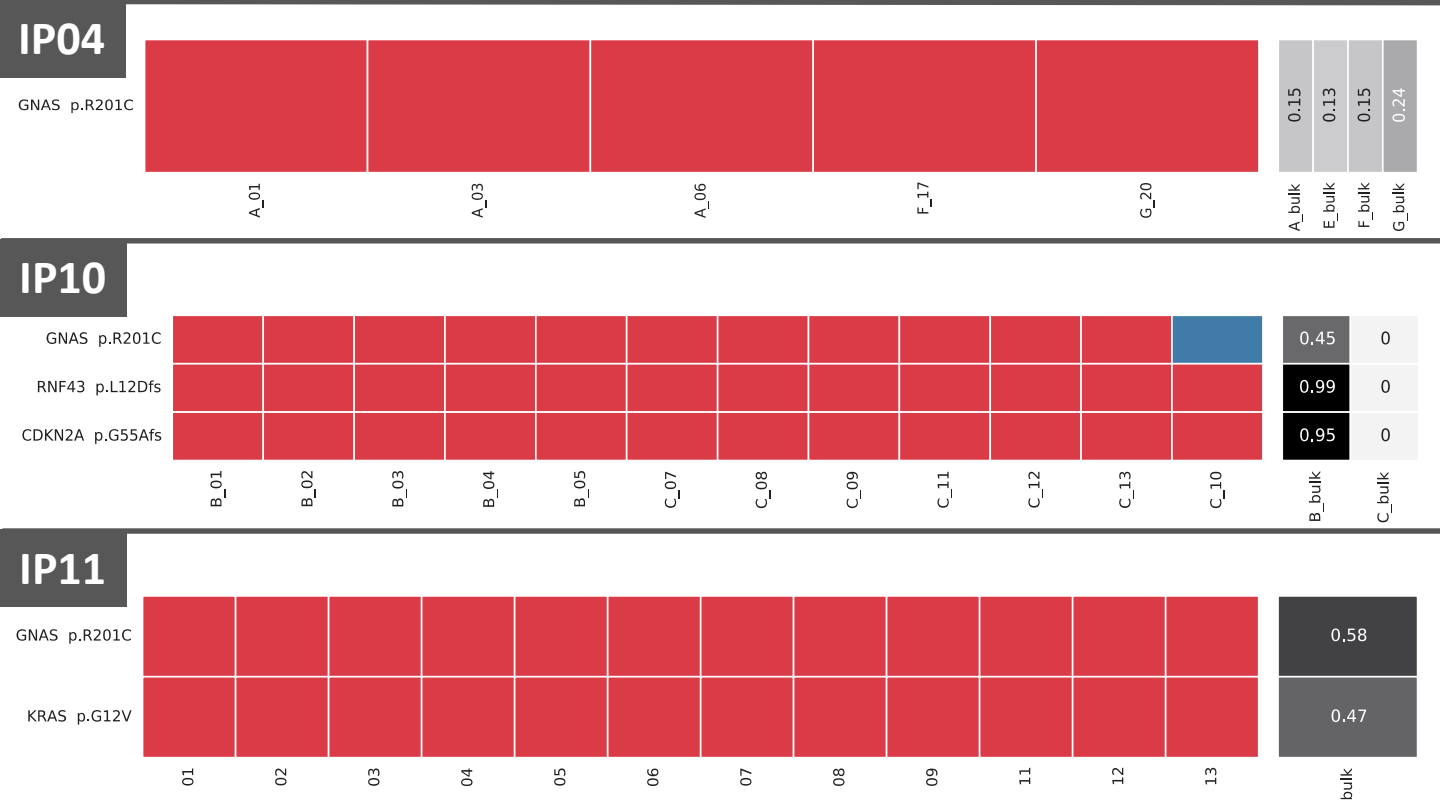

**Supplementary Figure S3. Thresholds applied for determining mutation status. A.** We started with BAM files for each IPMN sample (either from a *s*ingle *c*ell, *T*umor bulk, or *N*ormal bulk). For each of the single cell BAMs, variant calling was done with GATK HaplotypeCaller (independently for each one). Next, at the site of each called mutation in the single cells, BAMs for each single cell, tumor bulk and normal bulk were interrogated with SAMtools mpileup (default parameters). **B.** The output of this step was the variant allele frequency (VAF) and alternative read count at each site in each BAM. **C.** Bulk normal samples were used to identify germline mutations, and these were filtered out. **D.** Single cell mutation profile matrices were constructed. Each row represented a mutation and each column represented a single cell. Values of 0, 1, or 0.5 were assigned depending on whether the mutation was present, absent, or if its status was indeterminate, respectively, in the cell. **E.** Mutations occurring in only one single cell, and not in the bulk, were removed as likely false positives. We reasoned that the same false positive mutation, was unlikely to occur independently in two single cells or one single cell and one bulk sample, and such mutations were retained for further analysis. **F.** An iterative imputation algorithm was used to re-classify indeterminate sites.




**Supplementary Figure S4. Somatic mutation calls in single IPMN cells before and after imputation.** Somatic mutations are presented in heatmaps where each row represents a mutation and each column represents a single cell. Single cells are designated by their tissue section letter, if multiple sections were analyzed, and a single cell sample number. The colors indicate the mutation calls before (top) and after (bottom) imputation, with red indicating mutant, blue indicating wild-type, and black indicating indeterminate status of the mutation in the single cell sample. Our iterative procedure clustered single cell samples and mutations and imputed undetermined mutation status in a cell from a cluster of the most genetically similar cells in the tumor. IP20 is not shown, as all mutations in this IPMNs were false-positives by our criteria.




**Supplementary Figure S5. Outlier analysis of single cell ADO data.** Histogram distribution (top) and box plot (bottom) of single cell allelic dropout rate. Blue vertical lines mark the thresholds for outlier identification. Individual outliers are shown as grey diamonds on the boxplot.




**Supplementary Figure S6. Somatic mutations identified in single cells from IP04, IP10, IP11.** Somatic mutations are presented in heatmaps where each row represents a mutation and each column represents a single cell. Single cells are designated by their tissue section letter, if multiple sections were analyzed, and a single cell sample number. Cells and mutations were clustered with Euclidean distance bi-clustering. The colors indicate the mutation calls after imputation, with red indicating mutant and blue indicating wild-type. Variant allele frequencies of the identified mutations in bulk samples from each section are indicated on the right.